

# Learning Methods for Dynamic Topic Modeling in Automated Behaviour Analysis

Olga Isupova, Danil Kuzin, and Lyudmila Mihaylova, *Senior Member, IEEE*

**Abstract**—Semi-supervised and unsupervised systems provide operators with invaluable support and can tremendously reduce the operators' load. In the light of the necessity to process large volumes of video data and provide autonomous decisions, this work proposes new learning algorithms for activity analysis in video. The activities and behaviours are described by a dynamic topic model. Two novel learning algorithms based on the expectation maximisation approach and variational Bayes inference are proposed. Theoretical derivations of the posterior of model parameters are given. The designed learning algorithms are compared with the Gibbs sampling inference scheme introduced earlier in the literature. A detailed comparison of the learning algorithms is presented on real video data. We also propose an anomaly localisation procedure, elegantly embedded in the topic modeling framework. It is shown that the developed learning algorithms can achieve 95% success rate. The proposed framework can be applied to a number of areas, including transportation systems, security and surveillance.

**Index Terms**—behaviour analysis, unsupervised learning, learning dynamic topic models, variational Bayesian approach, expectation maximisation, video analytics

## I. INTRODUCTION

**B**EHAVIOUR analysis is an important area in intelligent video surveillance, where abnormal behaviour detection is a difficult problem. One of the challenges in this field is informality of problem formulation. Due to the broad scope of applications and desired objectives there is no unique way in which normal or abnormal behaviour can be described. In general the objective is to detect unusual events and inform in due course a human operator about them.

This paper considers a probabilistic framework for anomaly detection, where less probable events are labelled as abnormal. We propose two learning algorithms and an anomaly localisation procedure for spatial detection of abnormal behaviours.

### A. Related work

There is a wealth of methods for abnormal behaviour detection, for example, pattern-based methods [1]–[4]. The pattern-based methods extract explicit patterns from data and use them for decision making as behaviour templates. In [1] the sum of the visual features of a reference frame is treated as a normal behaviour template. Another common approach for representing normal templates is using clusters of visual features [2]–[4].

O.Isupova, D.Kuzin, L.Mihaylova are with the Department of Automatic Control and Systems Engineering, The University of Sheffield, Sheffield, UK e-mail: o.isupova@sheffield.ac.uk, dkuzin1@sheffield.ac.uk, l.s.mihaylova@sheffield.ac.uk

In the testing stage new observations are compared with the extracted patterns. The comparison is based on some similarity measure between observations, e.g., the Jensen-Shannon [5] divergence in [6] or the Z-score value in [3], [4]. If the distance between the new observation and any of the normal patterns is larger than a threshold then the observation is classified as abnormal.

Abnormal behaviour detection can be considered as a classification problem. It is difficult in advance to collect and label all kind of abnormalities. Therefore only one class label can be expected and one-class classifiers are applied to abnormal behaviour detection: e.g., a one-class Support Vector Machine [7], a support vector data description algorithm [8], a neural network approach [9], a level set method [10] for normal data boundary determination [11].

Another class of methods rely on the estimation of probability distributions of the visual data. These estimated distributions are then used in the decision making process. Different kinds of probability estimation algorithms are proposed in the literature, e.g., based on non-parametric sample histograms [12], Gaussian distribution modelling [13], [14]. Spatio-temporal motion data dependency is modelled as a coupled Hidden Markov Model in [15]. Auto-regressive process modelling based on self-organised maps is proposed in [16].

An efficient approach is to seek for feature sets that tend to appear together. These feature sets form the typical activities or behaviours in the scene. Topic modeling [17], [18] is an approach to find such kinds of statistical regularities in a form of probability distributions. The approach can be applied for abnormal behaviour detection, e.g., [19]–[21]. A number of variations of the conventional topic models for abnormal behaviour detection have been recently proposed: clustering of activities distributions [22], modelling temporal dependency among activities [23]–[25], a continuous model for an object velocity [26].

Within the probabilistic modelling approach [13]–[15], [19], [20], [22], [26] the decision about abnormality is mainly made by computing likelihood of a new observation. The comparison of the different abnormality measures based on the likelihood estimation is provided in [21].

Topic modeling is originally developed for text mining [17], [18]. It aims to find latent variables called “*topics*” given the collection of unlabelled text *documents* consisted of *words*. In probabilistic topic modeling documents are represented as a mixture of topics, where each topic is assumed to be a distribution over words.

There are two main types of topic models: Probabilistic Latent Semantic Analysis (PLSA) [17] and Latent Dirichlet

Allocation (LDA) [18]. The former considers the problem from the frequentist perspective while the later studies it within the Bayesian approach. The main learning techniques proposed for these models include maximum likelihood estimation via the Expectation-Maximisation (EM) algorithm [17], variational Bayes inference [18], Gibbs sampling [27], Maximum a Posteriori (MAP) estimation [28].

### B. Contributions

In this paper inspired by ideas from [23] we propose an unsupervised learning framework based on a Markov Clustering Topic Model for behaviour analysis and anomaly detection. It groups possible topic mixtures of visual documents and forms a Markov chain for the groups.

The key contributions of this work consist in developing new learning algorithms, namely MAP estimation using the EM-algorithm and variational Bayes inference for the Markov Clustering Topic Model (MCTM), and in proposing an anomaly localisation procedure that follows concepts of probabilistic topic modeling. We derive the likelihood expressions as a normality measure of newly observed data. The developed learning algorithms are compared with the Gibbs sampling scheme proposed in [23]. A comprehensive analysis of the algorithms is presented over real video sequences. The empirical results show that the proposed methods provide more accurate results than the Gibbs sampling scheme in terms of anomaly detection performance.

Our preliminary results with the EM-algorithm for behaviour analysis are published in [29]. In contrast to [29] we now consider a fully Bayesian framework, where we propose the EM-algorithm for MAP estimates rather than the maximum likelihood ones. We also propose here a novel learning algorithm based on variational Bayes inference and a novel anomaly localisation procedure. The experiments are performed on more challenging datasets in comparison to [29].

The rest of the paper is organised as follows. Section II describes the overall structure of visual documents and visual words. Section III introduces the dynamic topic model. The new learning algorithms are presented in Section IV, where the proposed MAP estimation via the EM-algorithm and variational Bayes algorithm are introduced first and then the Gibbs sampling scheme is reviewed. The methods are given with a detailed discussion about their similarities and differences. The anomaly detection procedure is presented in Section V. The learning algorithms are evaluated with real data in Section VI and Section VII concludes the paper.

## II. VIDEO ANALYTICS WITHIN THE TOPIC MODELING FRAMEWORK

Video analytics tasks can be formulated within the framework of topics modeling. This requires a definition of visual documents and visual words, e.g., as in [22]–[25]. The whole video sequence is divided into non-overlapping short clips. These clips are treated as visual documents. Each frame is divided next into grid cells of pixels. Motion detection is applied for each of the cells. The cells where motion is detected are called moving cells. For each of the moving cells

the motion direction is determined. This direction is further quantised into four dominant ones - up, left, down, right (see Figure 1). The position of the moving cell and the quantised direction of its motion form a visual word.

Each of the visual documents is then represented as a sequence of visual words' IDs, where IDs are obtained by some ordering of a set of unique words. This discrete representation of the input data can be processed by topic modeling methods.

## III. THE MARKOV CHAIN TOPIC MODEL FOR BEHAVIOURAL ANALYSIS

### A. Motivation

In topic modeling there are two main kinds of distributions — the distributions over words, which correspond to topics, and the distributions over topics, which characterise the documents. The relationship between documents and words is then represented via latent low-dimensional entities called topics. Having only an unlabelled collection of documents, topic modeling methods restore a hidden structure of data, i.e., the distributions over words and the distributions over topics.

Consider a set of distributions over topics and a topic distribution for each document is chosen from this set. If the cardinality of the set of distributions over topics is less than the number of documents then documents are clustered into groups, having the same topic distribution within a group. A unique distribution over topics is called a *behaviour* in this work. Therefore each document corresponds to one behaviour. In topic modeling a document is fully described by a corresponding distribution over topics, which means in this case a document is fully described by a corresponding behaviour.

There are a number of applications where we can observe documents clustered into groups with the same distribution over topics. Let us consider some examples from video analytics where a visual word corresponds to a motion within a tiny cell. As topics represent words that statistically often appear together, in video analytics applications topics define some motion patterns in local areas.

Let us consider a road junction regulated by traffic lights. A general motion on the junction is the same with the same traffic light regime. Therefore the documents associated to the same traffic light regimes have the same distributions over topics, i.e., they correspond to the same behaviours.

Another example is a video stream generated by a CCTV camera from a train station. Here it is also possible to distinguish several types of general motion within the camera scene: getting off and on a train and waiting for it. These types of motion correspond to behaviours, where the different visual documents showing different instances of the same behaviour have very similar motion structures, i.e., the same topic distribution.

Each action in real life lasts for some time, e.g., traffic light regime stays the same and people get on and off a train for several seconds. Moreover, often these different types of motion or behaviours follow a cycle and their changes occur in some order. These insights motivate to model a sequence of behaviours as a Markov chain, so that the behaviours remain



Figure 1. Structure of the visual feature extraction: from an input frame (on the left) a map of local motions is calculated (in the centre). The motion is quantised into four directions to get the feature representation (on the right).

the same during some documents and change in a predefined order. The model that has these described properties is called a Markov Clustering Topic Model (MCTM) in [23]. The next section formally formulates the model.

### B. Model formulation

This section starts from the introduction of the main notations used through the paper. Denote by  $\mathcal{X}$  the vocabulary of all visual words, by  $\mathcal{Y}$  the set of all topics and by  $\mathcal{Z}$  the set of all behaviours,  $x$ ,  $y$  and  $z$  are used to denote elements from these sets respectively. When an additional element of a set is required it denotes with a prime, e.g.  $z'$  is another element from  $\mathcal{Z}$ .

Let  $\mathbf{x}_t = \{x_{i,t}\}_{i=1}^{N_t}$  denote a set of words for the document  $t$ , where  $N_t$  is the length of the document  $t$ . Let  $\mathbf{x}_{1:T_{tr}} = \{\mathbf{x}_t\}_{t=1}^{T_{tr}}$  denote a set of all words for the whole dataset, where  $T_{tr}$  is the number of documents in the dataset. Similarly, denote by  $\mathbf{y}_t = \{y_{i,t}\}_{i=1}^{N_t}$  and  $\mathbf{y}_{1:T_{tr}} = \{\mathbf{y}_t\}_{t=1}^{T_{tr}}$  a set of topics for the document  $t$  and a set of all topics for the whole dataset respectively. Let  $\mathbf{z}_{1:T_{tr}} = \{\mathbf{z}_t\}_{t=1}^{T_{tr}}$  denote a set of all behaviours for all documents.

Note that  $x$ ,  $y$  and  $z$  without subscript denote possible values for a word, topic and behaviour from  $\mathcal{X}$ ,  $\mathcal{Y}$  and  $\mathcal{Z}$  respectively, while the symbols with subscript denote word, topic and behaviour assignments in particular places in a dataset.

Here,  $\Phi$  denotes a matrix corresponding to the distributions over words given the topics,  $\Theta$  denotes a matrix corresponding to the distributions over topics given behaviours. For a Markov chain of behaviours a vector  $\pi$  for a behaviour distribution for the first document and a matrix  $\Xi$  for transition probability distributions between the behaviours are introduced:

$$\begin{aligned}\Phi &= \{\phi_{x,y}\}_{x \in \mathcal{X}, y \in \mathcal{Y}}, & \phi_{x,y} &= p(x|y), & \phi_y &= \{\phi_{x,y}\}_{x \in \mathcal{X}}; \\ \Theta &= \{\theta_{y,z}\}_{y \in \mathcal{Y}, z \in \mathcal{Z}}, & \theta_{y,z} &= p(y|z), & \theta_z &= \{\theta_{y,z}\}_{y \in \mathcal{Y}}; \\ \pi &= \{\pi_z\}_{z \in \mathcal{Z}}, & \pi_z &= p(z); \\ \Xi &= \{\xi_{z',z}\}_{z',z \in \mathcal{Z}}, & \xi_{z',z} &= p(z'|z), & \xi_z &= \{\xi_{z',z}\}_{z' \in \mathcal{Z}},\end{aligned}$$

where the matrices  $\Phi$ ,  $\Theta$  and  $\Xi$  and the vector  $\pi$  are formed as follows. An element of a matrix on the  $i$ -th row and  $j$ -th column is a probability of the  $i$ -th element given the  $j$ -th

one, e.g.,  $\phi_{x,y}$  is a probability of the word  $x$  in the topic  $y$ . The columns of the matrices are then form distributions for a corresponding elements, e.g.,  $\theta_z$  is a distribution over topics for the behaviour  $z$ . Elements of the vector  $\pi$  are probabilities of behaviours to be chosen by the first document. All these distributions are categorical.

The introduced distributions form a set

$$\Omega = \{\Phi, \Theta, \pi, \Xi\} \quad (1)$$

of model parameters and they are estimated during a learning procedure.

Prior distributions are imposed to all the parameters. Conjugate Dirichlet distributions are used:

$$\begin{aligned}\phi_y &\sim \text{Dir}(\phi_y|\beta), & \forall y \in \mathcal{Y}; \\ \theta_z &\sim \text{Dir}(\theta_z|\alpha), & \forall z \in \mathcal{Z}; \\ \pi &\sim \text{Dir}(\pi|\eta); \\ \xi_z &\sim \text{Dir}(\xi_z|\gamma), & \forall z \in \mathcal{Z},\end{aligned}$$

where  $\text{Dir}(\cdot)$  denotes a Dirichlet distribution and  $\beta$ ,  $\alpha$ ,  $\eta$  and  $\gamma$  are the corresponding hyperparameters. As topics and behaviours are not known a priori and will be specified via the learning procedure, it is impossible to distinguish two topics or two behaviours in advance. This is the reason why all the prior distributions are the same for all topics and all behaviours.

The generative process for the model is as follows. All the parameters are drawn from the corresponding prior Dirichlet distributions. At each time moment  $t$  a behaviour  $z_t$  is chosen first for a visual document. The behaviour is sampled using the matrix  $\Xi$  according to the behaviour chosen for the previous document. For the first document the behaviour is sampled using the vector  $\pi$ .

Once the behaviour is selected, the procedure of choosing visual words repeats for the number of times equal to the length of the current document  $N_t$ . The procedure consists of two steps — sampling a topic  $y_{i,t}$  using the matrix  $\Theta$  according to the chosen behaviour  $z_t$  followed by sampling a word  $x_{i,t}$  using the matrix  $\Phi$  according to the chosen topic  $y_{i,t}$  for each token  $i \in \{1, \dots, N_t\}$ , where a token is a particular place inside a document where a word is assigned.

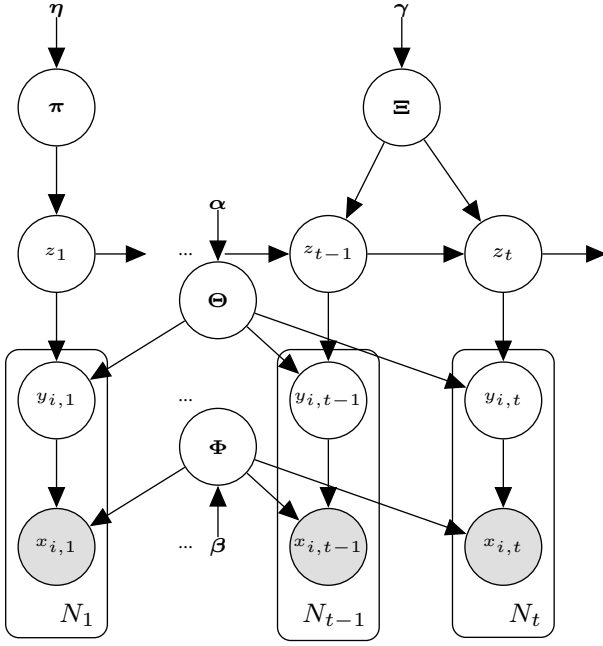


Figure 2. Graphical representation of the Markov Chain Topic Model.

The generative process is summarised in Algorithm III.1. The graphical model, showing the relationships between the variables, can be found in Figure 2.

The full likelihood of the observed variables  $\mathbf{x}_{1:T_{tr}}$ , the hidden variables  $\mathbf{y}_{1:T_{tr}}$  and  $\mathbf{z}_{1:T_{tr}}$  and the set of parameters  $\Omega$  can be written then as follows:

$$\begin{aligned}
 & p(\mathbf{x}_{1:T_{tr}}, \mathbf{y}_{1:T_{tr}}, \mathbf{z}_{1:T_{tr}}, \Omega | \beta, \alpha, \eta, \gamma) = \\
 & \underbrace{p(\pi | \eta) p(\Xi | \gamma) p(\Theta | \alpha) p(\Phi | \beta)}_{\text{Priors}} \times \\
 & \underbrace{p(z_1 | \pi) \left[ \prod_{t=2}^{T_{tr}} p(z_t | z_{t-1}, \Xi) \right] \prod_{t=1}^{T_{tr}} \prod_{i=1}^{N_t} p(x_{i,t} | y_{i,t}, \Phi) p(y_{i,t} | z_t, \Theta)}_{\text{Likelihood}}
 \end{aligned} \quad (2)$$

In [23] Gibbs sampling is implemented for parameters learning in the MCTM. We propose two new learning algorithms: based on an EM-algorithm for the MAP estimates of the parameters and based on variational Bayes inference to estimate posterior distributions of the parameters. We introduce the proposed learning algorithms below and briefly review the Gibbs sampling scheme.

#### IV. PARAMETERS LEARNING

##### A. Learning: EM-algorithm scheme

We propose a learning algorithm for MAP estimates of the parameters based on the Expectation-Maximisation algorithm [30]. The algorithm consists of repeating E and M-steps.

<sup>1</sup>Here,  $Cat(\cdot | \mathbf{v})$  denotes a categorical distribution, where components of a vector  $\mathbf{v}$  are probabilities of a discrete random variable to take one of possible values.

---

##### Algorithm III.1 The generative process for the MCTM

---

**Require:** The number of clips –  $T$ , the length of each clip –

$N_t \forall t = \{1, \dots, T\}$ , the hyperparameters –  $\beta, \alpha, \eta, \gamma$ ;

**Ensure:** The dataset  $\mathbf{x}_{1:T} = \{x_{1,1}, \dots, x_{i,t}, \dots, x_{N_T,T}\}$ ;

1: **for all**  $y \in \mathcal{Y}$  **do**

2: draw a word distribution for the topic  $y$ :

$$\phi_y \sim Dir(\phi_y | \beta);$$

3: **for all**  $z \in \mathcal{Z}$  **do**

4: draw a topic distribution for behaviour  $z$ :

$$\theta_z \sim Dir(\theta_z | \alpha);$$

5: draw a transition distribution for behaviour  $z$ :

$$\xi_z \sim Dir(\xi_z | \gamma);$$

6: draw a behaviour probability distribution for the initial document

$$\pi \sim Dir(\pi | \eta);$$

7: **for all**  $t \in \{1, \dots, T\}$  **do**

8: **if**  $t = 1$  **then**

9: draw a behaviour for the document from the initial distribution:  $z_t \sim Cat(z_t | \pi)^1$ ;

10: **else**

11: draw a behaviour for the document based on the behaviour of the previous document:  $z_t \sim Cat(z_t | \xi_{z_{t-1}})$ ;

12: **for all**  $i \in \{1, \dots, N_t\}$  **do**

13: draw a topic for the token  $i$  based on the chosen behaviour:  $y_{i,t} \sim Cat(y_{i,t} | \theta_{z_t})$ ;

14: draw a visual word for the token  $i$  based on the chosen topic:  $x_{i,t} \sim Cat(x_{i,t} | \phi_{y_{i,t}})$ ;

---

Conventionally, the EM-algorithm is applied to get maximum likelihood estimates. In that case the M-step is:

$$Q(\Omega, \Omega^{\text{old}}) \longrightarrow \max_{\Omega}, \quad (3)$$

where  $\Omega^{\text{old}}$  denotes the set of parameters obtained at the previous iteration and  $Q(\Omega, \Omega^{\text{old}})$  is the expected logarithm of the full likelihood function of the observed and hidden variables:

$$\begin{aligned}
 & Q(\Omega, \Omega^{\text{old}}) = \\
 & \mathbb{E}_{p(\mathbf{y}_{1:T_{tr}}, \mathbf{z}_{1:T_{tr}} | \mathbf{x}_{1:T_{tr}}, \Omega^{\text{old}})} \log p(\mathbf{x}_{1:T_{tr}}, \mathbf{y}_{1:T_{tr}}, \mathbf{z}_{1:T_{tr}} | \Omega). \quad (4)
 \end{aligned}$$

The subscript of the expectation sign means the distribution with respect to which the expectation is calculated. During the E-step the posterior distribution of the hidden variables is estimated given the current estimates of the parameters.

In this paper the EM-algorithm is applied to get MAP estimates instead of traditional maximum likelihood ones. The M-step is modified in this case as:

$$Q(\Omega, \Omega^{\text{old}}) + \log p(\Omega | \beta, \alpha, \eta, \gamma) \longrightarrow \max_{\Omega}, \quad (5)$$

where  $p(\Omega | \beta, \alpha, \eta, \gamma)$  is the prior distribution of the parameters.

As the hidden variables are discrete, the expectation converts to a sum of all possible values for the whole set of the hidden variables  $\{\mathbf{y}_{1:T_{tr}}, \mathbf{z}_{1:T_{tr}}\}$ . The substitution of the likelihood expression from (2) into (5) allows to marginalise some hidden variables from the sum. The remaining distributions that are required for computing the  $Q$ -function are follows:

- $p(z_1 = z | \mathbf{x}_{1:T_{tr}}, \mathbf{\Omega}^{\text{old}})$  — the posterior distribution of a behaviour for the first document;
- $p(z_t = z', z_{t-1} = z | \mathbf{x}_{1:T_{tr}}, \mathbf{\Omega}^{\text{old}})$  — the posterior distribution of two behaviours for successive documents;
- $p(y_{i,t} = y | \mathbf{x}_{1:T_{tr}}, \mathbf{\Omega}^{\text{old}})$  — the posterior distribution of a topic assignment for a given token;
- $p(y_{i,t} = y, z_t = z | \mathbf{x}_{1:T_{tr}}, \mathbf{\Omega}^{\text{old}})$  — the joint posterior distribution of a topic and behaviour assignments for a given token.

With the fixed current values for these posterior distributions the estimates of the parameters that maximise the required functional of the M-step (5) can be computed as:

$$\hat{\phi}_{x,y}^{\text{EM}} = \frac{(\beta_x + \hat{n}_{x,y}^{\text{EM}} - 1)_+}{\sum_{x' \in \mathcal{X}} (\beta_{x'} + \hat{n}_{x',y}^{\text{EM}} - 1)_+}, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}; \quad (6)$$

$$\hat{\theta}_{y,z}^{\text{EM}} = \frac{(\alpha_y + \hat{n}_{y,z}^{\text{EM}} - 1)_+}{\sum_{y' \in \mathcal{Y}} (\alpha_{y'} + \hat{n}_{y',z}^{\text{EM}} - 1)_+}, \quad \forall y \in \mathcal{Y}, z \in \mathcal{Z}; \quad (7)$$

$$\hat{\xi}_{z',z}^{\text{EM}} = \frac{(\gamma_{z'} + \hat{n}_{z',z}^{\text{EM}} - 1)_+}{\sum_{z \in \mathcal{Z}} (\gamma_z + \hat{n}_{z,z}^{\text{EM}} - 1)_+}, \quad \forall z', z \in \mathcal{Z}; \quad (8)$$

$$\hat{\pi}_z^{\text{EM}} = \frac{(\eta_z + \hat{n}_z^{\text{EM}} - 1)_+}{\sum_{z' \in \mathcal{Z}} (\eta_{z'} + \hat{n}_{z'}^{\text{EM}} - 1)_+}, \quad \forall z \in \mathcal{Z}, \quad (9)$$

where  $(a)_+ \stackrel{\text{def}}{=} \max(a, 0)$  [31];  $\beta_x$ ,  $\alpha_y$  and  $\gamma_{z'}$  are the elements of the hyperparameter vectors  $\beta$ ,  $\alpha$  and  $\gamma$  respectively, and:

- $\hat{n}_{x,y}^{\text{EM}} = \sum_{t=1}^T \sum_{i=1}^{N_t} p(y_{i,t} = y | \mathbf{x}_{1:T}, \mathbf{\Omega}^{\text{old}}) \mathbb{I}(x_{i,t} = x)$  — the expected number of times, when the word  $x$  is associated to the topic  $y$ , where  $\mathbb{I}(\cdot)$  is the indicator function;
- $\hat{n}_{y,z}^{\text{EM}} = \sum_{t=1}^T \sum_{i=1}^{N_t} p(y_{i,t} = y, z_t = z | \mathbf{x}_{1:T}, \mathbf{\Omega}^{\text{old}})$  — the expected number of times, when the topic  $y$  is associated to the behaviour  $z$ ;
- $\hat{n}_z^{\text{EM}} = p(z_1 = z | \mathbf{x}_{1:T}, \mathbf{\Omega}^{\text{old}})$  — the “expected number of times”, when the behaviour  $z$  is associated to the first document, in this case the “expected number” is just a probability, the notation is used for the similarity with the rest of the parameters;
- $\hat{n}_{z',z}^{\text{EM}} = \sum_{t=2}^T p(z_t = z', z_{t-1} = z | \mathbf{x}_{1:T}, \mathbf{\Omega}^{\text{old}})$  — the expected number of times, when the behaviour  $z$  is followed by the behaviour  $z'$ .

During the E-step with the fixed current estimates of the parameters  $\mathbf{\Omega}^{\text{old}}$ , the updated values for the posterior distributions of the hidden variables should be computed. The derivation of the updated formulae for these distributions is similar to the Baum-Welch forward-backward algorithm [32], where the EM-algorithm is applied to the maximum likelihood estimates

for a Hidden Markov Model (HMM). This similarity appears because the generative model can be viewed as extension of a HMM.

For effective computation of the required posterior distributions the additional variables  $\hat{\alpha}_z(t)$  and  $\hat{\beta}_z(t)$  are introduced. A dynamic programming technique is applied for computation of these variables. Having the updated values for  $\hat{\alpha}_z(t)$  and  $\hat{\beta}_z(t)$  one can update the required posterior distributions of the hidden variables. The E-step is then formulated as follows (for simplification of notation the superscript “old” for the parameters variables is omitted inside the formulae):

$$\begin{cases} \hat{\alpha}_z(t) = \prod_{i=1}^{N_t} \sum_{y \in \mathcal{Y}} \phi_{x_{i,t}, y} \theta_{y,z} \times \\ \sum_{z' \in \mathcal{Z}} \hat{\alpha}_{z'}(t-1) \xi_{z,z'}, \text{ if } t \geq 2, \forall z \in \mathcal{Z}; \end{cases} \quad (10)$$

$$\begin{cases} \hat{\alpha}_z(1) = \pi_z \prod_{i=1}^{N_1} \sum_{y \in \mathcal{Y}} \phi_{x_{i,1}, y} \theta_{y,z}, \forall z \in \mathcal{Z}; \\ \hat{\beta}_z(t) = \sum_{z' \in \mathcal{Z}} \hat{\beta}_{z'}(t+1) \xi_{z',z} \times \\ \prod_{i=1}^{N_{t+1}} \sum_{y \in \mathcal{Y}} \phi_{x_{i,t+1}, y} \theta_{y,z'}, \text{ if } t \leq T-1, \forall z \in \mathcal{Z}; \\ \hat{\beta}_z(T) = 1, \forall z \in \mathcal{Z}; \end{cases} \quad (11)$$

$$K = \sum_{z \in \mathcal{Z}} \hat{\alpha}_z(1) \hat{\beta}_z(1); \quad (12)$$

$$p(z_1 | \mathbf{x}_{1:T}, \mathbf{\Omega}^{\text{old}}) = \frac{\hat{\alpha}_{z_1}(1) \hat{\beta}_{z_1}(1)}{K}; \quad (13)$$

$$p(z_t, z_{t-1} | \mathbf{x}_{1:T}, \mathbf{\Omega}^{\text{old}}) = \frac{\hat{\alpha}_{z_{t-1}}(t-1) \hat{\beta}_{z_t}(t) \xi_{z_t, z_{t-1}} \times \prod_{i=1}^{N_t} \sum_{y \in \mathcal{Y}} \phi_{x_{i,t}, y} \theta_{y,z_t}}{K} \quad (14)$$

$$\begin{cases} p(y_{i,t}, z_t | \mathbf{x}_{1:T}, \mathbf{\Omega}^{\text{old}}) = \frac{1}{K} \phi_{x_{i,t}, y_{i,t}} \theta_{y_{i,t}, z_t} \hat{\beta}_{z_t}(t) \times \\ \sum_{z' \in \mathcal{Z}} \hat{\alpha}_{z'}(t-1) \xi_{z_t, z'} \times \\ \prod_{j=1}^{N_t} \sum_{y' \in \mathcal{Y}} \phi_{x_{j,t}, y'} \theta_{y', z_t}, \text{ if } t \geq 2; \\ p(y_{i,1}, z_1 | \mathbf{x}_{1:T}, \mathbf{\Omega}^{\text{old}}) = \frac{1}{K} \phi_{x_{i,1}, y_{i,1}} \theta_{y_{i,1}, z_1} \hat{\beta}_{z_1}(1) \times \\ \pi_{z_1} \prod_{j=1}^{N_1} \sum_{y' \in \mathcal{Y}} \phi_{x_{j,1}, y'} \theta_{y', z_1}; \end{cases} \quad (15)$$

$$p(y_{i,t} | \mathbf{x}_{1:T}, \mathbf{\Omega}^{\text{old}}) = \sum_{z \in \mathcal{Z}} p(y_{i,t}, z | \mathbf{x}_{1:T}, \mathbf{\Omega}^{\text{old}}), \quad (16)$$

where  $K$  is a normalisation constant for all the posterior distributions of the hidden variables.

Starting with some random initialisation of the parameters estimates, the EM-algorithm iterates the E and M-steps until convergence. The obtained estimates of the parameters are used for further analysis.

### B. Learning: Variational Bayes scheme

We also propose a learning algorithm based on the variational Bayes (VB) approach [33] to find approximated posterior distributions for both the hidden variables and the parameters.

In the VB inference scheme the true posterior distribution, in this case the distribution of the parameters and the hidden variables  $p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \mathbf{\Omega} | \mathbf{x}_{1:T}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ , is approximated with a factorised distribution —  $q(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \mathbf{\Omega})$ . The approximation is made to minimise the Kullback-Leibler divergence between the factorised distribution and true one. We factorise the distribution in order to separate the hidden variables and the parameters:

$$\hat{q}(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \mathbf{\Omega}) = \hat{q}(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}) \hat{q}(\mathbf{\Omega}) \stackrel{\text{def}}{=} \underset{\text{argmin}}{\text{KL}}(q(\mathbf{y}_{1:T}, \mathbf{z}_{1:T})q(\mathbf{\Omega}) || p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \mathbf{\Omega} | \mathbf{x}_{1:T}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta})), \quad (17)$$

where KL denotes the Kullback-Leibler divergence. The minimisation of the Kullback-Leibler divergence is equivalent to the maximisation of the evidence lower bound (ELBO). The maximisation is done by coordinate ascent [33].

During the update of the parameters the approximated distribution  $q(\mathbf{\Omega})$  it is further factorised:

$$q(\mathbf{\Omega}) = q(\boldsymbol{\pi})q(\boldsymbol{\Xi})q(\boldsymbol{\Theta})q(\boldsymbol{\Phi}). \quad (18)$$

Note that this factorisation is a corollary of our model and not an assumption.

The iterative process of updating the approximated distributions of the parameters and the hidden variables can be formulated as an EM-like algorithm, where during the E-step the approximated distributions of the hidden variables are updated and during the M-step the approximated distributions of the parameters are updated.

The M-like step is as follows:

$$\begin{cases} q(\boldsymbol{\Phi}) = \prod_{y \in \mathcal{Y}} \text{Dir}(\phi_y; \tilde{\beta}_y), \\ \tilde{\beta}_{x,y} = \beta_x + \hat{n}_{x,y}^{\text{VB}}, \end{cases} \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}; \quad (19)$$

$$\begin{cases} q(\boldsymbol{\Theta}) = \prod_{z \in \mathcal{Z}} \text{Dir}(\theta_z; \tilde{\alpha}_z), \\ \tilde{\alpha}_{y,z} = \alpha_y + \hat{n}_{y,z}^{\text{VB}}, \end{cases} \quad \forall y \in \mathcal{Y}, z \in \mathcal{Z}; \quad (20)$$

$$\begin{cases} q(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}; \tilde{\eta}), \\ \tilde{\eta}_z = \eta_z + \hat{n}_z^{\text{VB}}, \end{cases} \quad \forall z \in \mathcal{Z}; \quad (21)$$

$$\begin{cases} q(\boldsymbol{\Xi}) = \prod_{z \in \mathcal{Z}} \text{Dir}(\boldsymbol{\xi}_z; \tilde{\gamma}_z), \\ \tilde{\gamma}_{z',z} = \gamma_{z'} + \hat{n}_{z',z}^{\text{VB}}, \end{cases} \quad \forall z', z \in \mathcal{Z}, \quad (22)$$

where  $\tilde{\beta}_y$ ,  $\tilde{\alpha}_z$ ,  $\tilde{\eta}$  and  $\tilde{\gamma}_z$  are updated hyperparameters of the corresponding posterior Dirichlet distributions and

- $\hat{n}_{x,y}^{\text{VB}} = \sum_{t=1}^T \sum_{i=1}^{N_t} \mathbb{I}(x_{i,t} = x) q(y_{i,t} = y)$  — the expected number of times, when the word  $x$  is associated with the topic  $y$ . Here and below the expected number is computed with respect to the approximated posterior distributions of the hidden variables;

- $\hat{n}_{y,z}^{\text{VB}} = \sum_{t=1}^T \sum_{i=1}^{N_t} q(y_{i,t} = y, z_t = z)$  — the expected number of times, when the topic  $y$  is associated with the behaviour  $z$ ;
- $\hat{n}_z^{\text{VB}} = q(z_1 = z)$  — the “expected number” of times, when the behaviour  $z$  is associated to the first document;
- $\hat{n}_{z',z}^{\text{VB}} = \sum_{t=2}^T q(z_t = z', z_{t-1} = z)$  — the expected number of times, when the behaviour  $z$  is followed by the behaviour  $z'$ .

The following additional variables are introduced for the E-like step:

$$\tilde{\pi}_z = \exp \left( \psi(\tilde{\eta}_z) - \psi \left( \sum_{z' \in \mathcal{Z}} \tilde{\eta}_{z'} \right) \right), \quad \forall z \in \mathcal{Z}; \quad (23)$$

$$\tilde{\xi}_{\tilde{z},z} = \exp \left( \psi(\tilde{\gamma}_{\tilde{z},z}) - \psi \left( \sum_{z' \in \mathcal{Z}} \tilde{\gamma}_{z',z} \right) \right), \quad \forall \tilde{z}, z \in \mathcal{Z}; \quad (24)$$

$$\tilde{\phi}_{x,y} = \exp \left( \psi(\tilde{\beta}_{x,y}) - \psi \left( \sum_{x' \in \mathcal{X}} \tilde{\beta}_{x',y} \right) \right), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}; \quad (25)$$

$$\tilde{\theta}_{y,z} = \exp \left( \psi(\tilde{\alpha}_{y,z}) - \psi \left( \sum_{y' \in \mathcal{Y}} \tilde{\alpha}_{y',z} \right) \right), \quad \forall y \in \mathcal{Y}, z \in \mathcal{Z}, \quad (26)$$

where  $\psi(\cdot)$  is the digamma function.

Using these additional notations the E-like step is formulated the same as the E-step of the EM-algorithm, replacing everywhere the estimates of the parameters with the corresponding tilde introduced notation and true posterior distributions of the hidden variables with the corresponding approximated ones in (10)-(16).

The point estimates of the parameters can be obtained by expected values of the posterior approximated distributions. An expected value for a Dirichlet distribution (a posterior distribution for all the parameters) is a normalised vector of hyperparameters. Using the expressions for the hyperparameters from (19) – (22), the final parameters estimates can be obtained by:

$$\hat{\phi}_{x,y}^{\text{VB}} = \frac{\beta_x + \hat{n}_{x,y}^{\text{VB}}}{\sum_{x' \in \mathcal{X}} (\beta_{x'} + \hat{n}_{x',y}^{\text{VB}})}, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}; \quad (27)$$

$$\hat{\theta}_{y,z}^{\text{VB}} = \frac{\alpha_y + \hat{n}_{y,z}^{\text{VB}}}{\sum_{y' \in \mathcal{Y}} (\alpha_{y'} + \hat{n}_{y',z}^{\text{VB}})}, \quad \forall y \in \mathcal{Y}, z \in \mathcal{Z}; \quad (28)$$

$$\hat{\xi}_{z',z}^{\text{VB}} = \frac{\gamma_{z'} + \hat{n}_{z',z}^{\text{VB}}}{\sum_{\tilde{z} \in \mathcal{Z}} (\gamma_{\tilde{z}} + \hat{n}_{\tilde{z},z}^{\text{VB}})}, \quad \forall z', z \in \mathcal{Z}; \quad (29)$$

$$\hat{\pi}_z^{\text{VB}} = \frac{\eta_z + \hat{n}_z^{\text{VB}}}{\sum_{z' \in \mathcal{Z}} (\eta_{z'} + \hat{n}_{z'}^{\text{VB}})}, \quad \forall z \in \mathcal{Z}. \quad (30)$$

### C. Learning: Gibbs sampling algorithm

In [23] the collapsed version of Gibbs sampling (GS) is used for parameter learning in the MCTM. The Markov chain is built to sample only the hidden variables  $y_{i,t}$  and  $z_t$ , while the parameters  $\Phi$ ,  $\Theta$  and  $\Xi$  are integrated out (note that the distribution for the initial behaviour choice  $\pi$  is not considered in [23]).

During the burn-in stage the hidden topic and behaviour assignments to each token in the dataset are drawn from the conditional distributions given all the remaining variables. Following the Markov Chain Monte Carlo framework it would draw samples from the posterior distribution  $p(\mathbf{y}_{1:T_{tr}}, \mathbf{z}_{1:T_{tr}} | \mathbf{x}_{1:T_{tr}}, \beta, \alpha, \eta, \gamma)$ . From the whole sample for  $\{\mathbf{y}_{1:T_{tr}}, \mathbf{z}_{1:T_{tr}}\}$  the parameters can be estimated by [27]:

$$\hat{\phi}_{x,y}^{\text{GS}} = \frac{\hat{n}_{x,y}^{\text{GS}} + \beta_x}{\sum_{x' \in \mathcal{X}} (\hat{n}_{x',y}^{\text{GS}} + \beta_{x'})}, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}; \quad (31)$$

$$\hat{\theta}_{y,z}^{\text{GS}} = \frac{\hat{n}_{y,z}^{\text{GS}} + \alpha_y}{\sum_{y' \in \mathcal{Y}} (\hat{n}_{y',z}^{\text{GS}} + \alpha_{y'})}, \quad \forall y \in \mathcal{Y}, z \in \mathcal{Z}; \quad (32)$$

$$\hat{\xi}_{z',z}^{\text{GS}} = \frac{\hat{n}_{z',z}^{\text{GS}} + \gamma_{z'}}{\sum_{z \in \mathcal{Z}} (\hat{n}_{z,z}^{\text{GS}} + \gamma_z)}, \quad \forall z', z \in \mathcal{Z}, \quad (33)$$

where  $\hat{n}_{x,y}^{\text{GS}}$  is the count for the number of times when the word  $x$  is associated with the topic  $y$ ,  $\hat{n}_{y,z}^{\text{GS}}$  is the count for the topic  $y$  and the behaviour  $z$  pair,  $\hat{n}_{z',z}^{\text{GS}}$  is the count for the number of times when the behaviour  $z$  is followed by the behaviour  $z'$ .

### D. Similarities and differences of the learning algorithms

The point parameter estimates for all three learning algorithms (6)–(9), (27)–(30) and (31)–(33) have a similar form. The EM-algorithm estimates differ up to the hyperparameters reassignment — adding one to all the hyperparameters in the VB or GS algorithms ends up with the same final equations for the parameters estimates in the EM-algorithm. We explore this in the experimental part. This “-1” term in the EM-algorithm formulae (6)–(8) occurs because it uses modes of the posterior distributions while the point estimates obtained by the VB and GS algorithms are means of the corresponding posterior distributions. For a Dirichlet distribution, which is a posterior distribution for all the parameters, mode and mean expressions differ by this “-1” term.

The main differences of the methods consist in the ways the counts  $n_{x,y}$ ,  $n_{y,z}$  and  $n_{z',z}$  are estimated. In the GS algorithm they are calculated by a single sample from the posterior distribution of the hidden variables  $p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T} | \mathbf{x}_{1:T}, \beta, \alpha, \eta, \gamma)$ . In the EM-algorithm the counts are computed as expected numbers of the corresponding events with respect to the posterior distributions of the hidden variables. In the VB algorithm the counts are computed in the same way as in the EM-algorithm up to replacing the true posterior distributions with the approximated ones.

Our observations for the dynamic topic model confirm the comparison results for the vanilla PLSA and LDA models provided in [34].

## V. ANOMALY DETECTION

This paper presents on-line anomaly detection with the MCTM in video streams. The decision making procedure is divided into two stages. At a learning stage the parameters are estimated using  $T_{tr}$  visual documents by one of the learning algorithms, presented in Section IV. After that during a testing stage a decision about abnormality of new upcoming testing documents is made comparing a marginal likelihood of each document with a threshold. The likelihood is computed using the parameters obtained during the learning stage. The threshold is a parameter of the method and can be set empirically, for example, to label 2% of the testing data as abnormal. This paper presents a comparison of the algorithms (Section VI) using the measure independent of threshold value selection.

We also propose an anomaly localisation procedure during the testing stage for those visual documents that are labelled as abnormal. This procedure is designed to provide spatial information about anomalies, while documents labelled as abnormal provide temporal detection. The following sections introduce both the anomaly detection procedure on a document level and the anomaly localisation procedure within a video frame.

### A. Abnormal documents detection

The marginal likelihood of a new visual document  $\mathbf{x}_{t+1}$  given all the previous data  $\mathbf{x}_{1:t}$  can be used as normality measure of the document [23]:

$$p(\mathbf{x}_{t+1} | \mathbf{x}_{1:t}) = \iiint p(\mathbf{x}_{t+1} | \mathbf{x}_{1:t}, \Phi, \Theta, \Xi) p(\Phi, \Theta, \Xi | \mathbf{x}_{1:t}) d\Phi d\Theta d\Xi. \quad (34)$$

If the likelihood value is small it means that the current document cannot be fitted to the learnt behaviours and topics, which represent typical motion patterns. Therefore this is an indication for an abnormal event in this document. The decision about abnormality of a document is then made by comparing the marginal likelihood of the document with the threshold.

In real world applications it is essential to detect anomalies as soon as possible. Hence an approximation of the integral in (34) is used for efficient computation. The first approximation is based on the assumption that the training dataset is representative for parameter learning that means that posterior probability of the parameters would not change if there is more observed data:

$$p(\Phi, \Theta, \Xi | \mathbf{x}_{1:t}) \approx p(\Phi, \Theta, \Xi | \mathbf{x}_{1:T_{tr}}) \quad \forall t \geq T_{tr}. \quad (35)$$

The marginal likelihood can be then approximated as

$$\iiint p(\mathbf{x}_{t+1} | \mathbf{x}_{1:t}, \Phi, \Theta, \Xi) p(\Phi, \Theta, \Xi | \mathbf{x}_{1:t}) d\Phi d\Theta d\Xi \approx \iiint p(\mathbf{x}_{t+1} | \mathbf{x}_{1:t}, \Phi, \Theta, \Xi) p(\Phi, \Theta, \Xi | \mathbf{x}_{1:T_{tr}}) d\Phi d\Theta d\Xi. \quad (36)$$

Depending on the algorithm used for learning the integral in (36) can be further approximated in different ways. We consider two types of approximation.



1) *Plug-in approximation:* We propose a plug-in approximation [35] of (36) that uses point estimates of the parameters:

$$\begin{aligned} & \iiint p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \Phi, \Theta, \Xi) p(\Phi, \Theta, \Xi|\mathbf{x}_{1:T_r}) d\Phi d\Theta d\Xi \approx \\ & \iiint p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \Phi, \Theta, \Xi) \delta_{\hat{\Phi}}(\Phi) \delta_{\hat{\Theta}}(\Theta) \delta_{\hat{\Xi}}(\Xi) d\Phi d\Theta d\Xi = \\ & p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \hat{\Phi}, \hat{\Theta}, \hat{\Xi}), \end{aligned} \quad (37)$$

where  $\delta_a(\cdot)$  is the delta-function with the centre in  $a$ ;  $\hat{\Phi}$ ,  $\hat{\Theta}$ ,  $\hat{\Xi}$  are point estimates of the parameters which can be computed by any of the considered learning algorithms using (6)–(8), (27)–(29) or (31)–(33).

The product and sum rules, the conditional independence equations from the generative model are then applied and the final formula for the plug-in approximation is follows:

$$\begin{aligned} p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}) & \approx p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \hat{\Phi}, \hat{\Theta}, \hat{\Xi}) = \\ & \sum_{z_t} \sum_{z_{t+1}} \left[ p(\mathbf{x}_{t+1}|z_{t+1}, \hat{\Phi}, \hat{\Theta}) \times \right. \\ & \left. p(z_{t+1}|z_t, \hat{\Xi}) p(z_t|\mathbf{x}_{1:t}, \hat{\Phi}, \hat{\Theta}, \hat{\Xi}) \right], \end{aligned} \quad (38)$$

where the predictive probability of the behaviour for the current document, given the observed data up to the current document, can be computed via the recursive formula:

$$\begin{aligned} p(z_t|\mathbf{x}_{1:t}, \hat{\Phi}, \hat{\Theta}, \hat{\Xi}) & = \\ & \sum_{z_{t-1}} \frac{p(\mathbf{x}_t|z_t, \hat{\Phi}, \hat{\Theta}) p(z_t|z_{t-1}, \hat{\Xi}) p(z_{t-1}|\mathbf{x}_{1:t-1}, \hat{\Phi}, \hat{\Theta}, \hat{\Xi})}{p(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \hat{\Phi}, \hat{\Theta}, \hat{\Xi})}. \end{aligned} \quad (39)$$

The point estimates can be computed for all three learning algorithms, therefore a normality measure based on the plug-in approximation of the marginal likelihood is applicable for all of them.

2) *Monte Carlo approximation:* If samples  $\{\Phi^s, \Theta^s, \Xi^s\}$  from the posterior distribution  $p(\Phi, \Theta, \Xi|\mathbf{x}_{1:T_r})$  of the parameters can be obtained, the integral (36) is further approximated by the Monte Carlo method:

$$\begin{aligned} & \iiint p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \Phi, \Theta, \Xi) p(\Phi, \Theta, \Xi|\mathbf{x}_{1:T_r}) d\Phi d\Theta d\Xi \approx \\ & \frac{1}{S} \sum_{s=1}^S p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \Phi^s, \Theta^s, \Xi^s), \end{aligned} \quad (40)$$

where  $S$  is the number of samples. These samples can be obtained (i) from the approximated posterior distributions  $q(\Phi)$ ,  $q(\Theta)$ , and  $q(\Xi)$  of the parameters, computed by the VB learning algorithm, or (ii) from the independent samples of the GS scheme. For the conditional likelihood  $p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}, \Phi^s, \Theta^s, \Xi^s)$  the formula (38) is valid.

Note that for the approximated posterior distribution of the parameters, i.e., the output of the VB learning algorithm, the integral (36) can be resolved analytically, but it would be computationally infeasible. This is the reason why the Monte Carlo approximation is used in this case.

Finally, in order to compare documents of different lengths the normalised likelihood is used as a normality measure  $s$ :

$$s(\mathbf{x}_{t+1}) = \frac{1}{N_{t+1}} p(\mathbf{x}_{t+1}|\mathbf{x}_{1:t}). \quad (41)$$

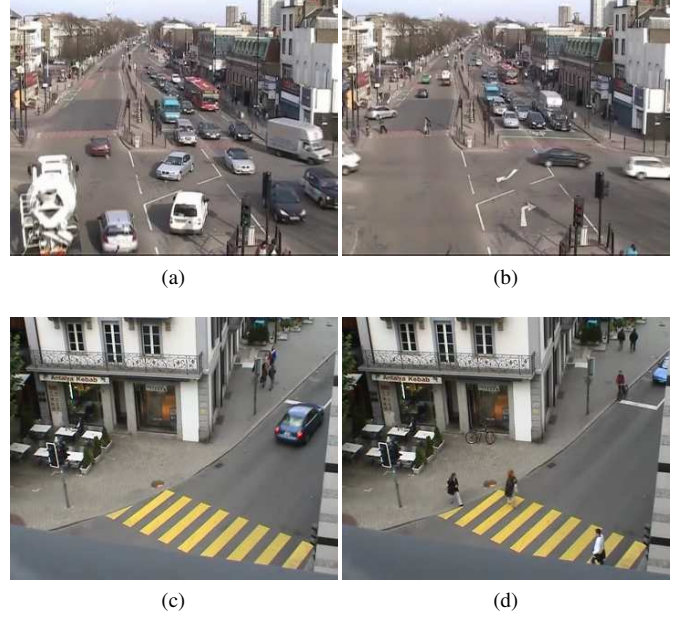


Figure 3. Sample frames of the real datasets. The top row presents two sample frames from the QMUL data, the bottom row presents two sample frames from the Idiap data.

## B. Localisation of anomalies

The topic modeling approach allows to compute a likelihood function not only of the whole document but of an individual word within the document too. Recall that the visual word contains the information about a location in the frame. We propose to use the location information from the least probable words (e.g., 10 words with the least likelihood values) to localise anomalies in the frame. Note, we do not require anything additional to a topic model, e.g., modelling regional information explicitly as in [36] or comparing a test document with training ones as in [37]. Instead, the proposed anomaly localisation procedure is general and can be applied in any topic modeling based method, where spatial information is encoded to visual words.

The marginal likelihood of a word can be computed in a similar way to the likelihood of the whole document. For the point estimates of the parameters and plug-in approximation of the integral it is:

$$p(x_{i,t+1}|\mathbf{x}_{1:t}) \approx p(x_{i,t+1}|\mathbf{x}_{1:t}, \hat{\Phi}, \hat{\Theta}, \hat{\Xi}). \quad (42)$$

For the samples from the posterior distributions of the parameters and the Monte Carlo integral approximation it is:

$$p(x_{i,t+1}|\mathbf{x}_{1:t}) \approx \frac{1}{S} \sum_{s=1}^S p(x_{i,t+1}|\mathbf{x}_{1:t}, \Phi^s, \Theta^s, \Xi^s). \quad (43)$$

## VI. PERFORMANCE VALIDATION

We compare the two proposed learning algorithms, based on EM and VB, with the GS algorithm, proposed in [23], on two real datasets.



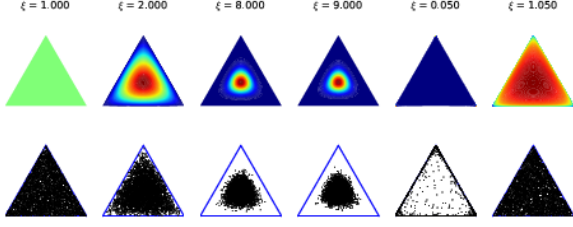


Figure 4. Dirichlet distributions with different symmetric parameters  $\xi$ . For the representation purposes the three-dimensional space is used. On the top row the colours correspond to the Dirichlet probability density function values in the area. On the bottom row there are samples generated from the corresponding density functions. The sample size is 5000.

### A. Setup

The performance of the algorithms is compared on the QMUL street intersection data [23] and Idiap traffic junction data [21]. Both datasets are 45-minutes video sequences, captured busy traffic road junctions, where we use a 5-minute video sequence as a training dataset and others as a testing one. The documents that have less than 20 visual words are discarded from consideration. In practice these documents can be classified to be normal by default as there is no enough information to make a decision. The frame size for both datasets is  $288 \times 360$ . Sample frames are presented in Figure 3.

The size of grid cells is set to  $8 \times 8$  pixels for spatial quantisation of the local motion for visual word determination. Non-overlapping clips with a one second length are treated as visual documents.

We also study the influence of the hyperparameters on the learning algorithms. In all the experiments we use the symmetric hyperparameters:  $\alpha = \{\alpha, \dots, \alpha\}$ ,  $\beta = \{\beta, \dots, \beta\}$ ,  $\gamma = \{\gamma, \dots, \gamma\}$  and  $\pi = \{\pi, \dots, \pi\}$ . The three groups of the hyperparameters settings are compared:  $\{\alpha = 1, \beta = 1, \gamma = 1, \pi = 1\}$  (referred as “prior type 1”),  $\{\alpha = 8, \beta = 0.05, \gamma = 1, \pi = 1\}$  (“prior type H”) and  $\{\alpha = 9, \beta = 1.05, \gamma = 2, \pi = 2\}$  (“prior type H+1”). Note that the first group corresponds to the case when in the EM-algorithm learning scheme the prior components are cancelled out, i.e., the MAP estimates in this case are equal to the maximum likelihood ones. The equations for the point estimates in the EM learning algorithm with the prior type H+1 of the hyperparameters settings are equal to the equations for the point estimates in the VB and GS learning algorithms with the prior type H of the settings. The corresponding Dirichlet distributions with all used parameters are presented in Figure 4.

Note, that parameter learning is an ill-posed problem in topic modeling [31]. This means there is no unique solution for parameter estimates. We use 20 Monte Carlo runs for all the learning algorithms with different random initialisations resulting with different solutions. The mean results among these runs are presented below for comparison.

All three algorithms are run with three different groups of hyperparameters settings. The number of topics and behaviours is set to 8 and 4, respectively for the QMUL dataset, 10 and 3 are used for the corresponding values for the Idiap dataset. The EM and VB algorithms are run for 100 iterations.

The GS algorithm is run for 500 burn-in iterations and 5 independent samples are taken with a 100 iterations delay after the burn-in period.

### B. Performance measure

Anomaly detection performance of the algorithms depends on threshold selection. To make a fair comparison of the different learning algorithms we use a performance measure which is independent of threshold selection.

In binary classification the following measures [35] are used:

- TP — true positive, a number of documents which are correctly detected as positive (abnormal in our case);
- TN — true negative, a number of documents which are correctly detected as negative (normal in our case);
- FP — false positive, a number of documents which are incorrectly detected as positive, when they are negative;
- FN — false negative, a number of documents which are incorrectly detected as negative, when they are positive;
- precision =  $\frac{TP}{TP + FP}$  — a fraction of correct detections among all documents labelled as abnormal by an algorithm;
- recall =  $\frac{TP}{TP + FN}$  — a fraction of correct detections among all truly abnormal documents.

The area under the precision-recall curve is used as a performance measure in this paper. This measure is more informative for detection of rare events than the popular area under the receiver operating characteristic (ROC) curve [35].

### C. Parameter learning

We visualise the learnt behaviours for the qualitative assessment of the proposed framework (Figures 5 and 6). For illustrative purposes we consider one run of the EM learning algorithm with the prior type H+1 of the hyperparameters settings.

The behaviours learnt for the QMUL data are shown in Figure 5 (for visualisation words representing 50% probability mass of a behaviour are used). One can notice that the algorithm correctly recognises the motion patterns in the data. The general motion of the scene follows a cycle: a vertical traffic flow (the first behaviour in Figure 5a), when cars move downward and upward on the road; left and right turns (the fourth behaviour in Figure 5d): some cars moving on the “vertical” road turn to the perpendicular road at the end of the vertical traffic flow; a left traffic flow (the second behaviour in Figure 5b), when cars move from right to left on the “horizontal” road; and a right traffic flow (the third behaviour in Figure 5c), when cars move from left to right on the “horizontal” road. Note, that the ordering numbers of behaviours correspond to their internal representation in the algorithm. The transition probability matrix  $\Xi$  is used to recognise the correct behaviours order in the data.

Figure 6 presents the behaviours learnt for the Idiap data. In this case the learnt behaviours have also a clear semantic meaning. The scene motion follows a cycle: a pedestrians flow

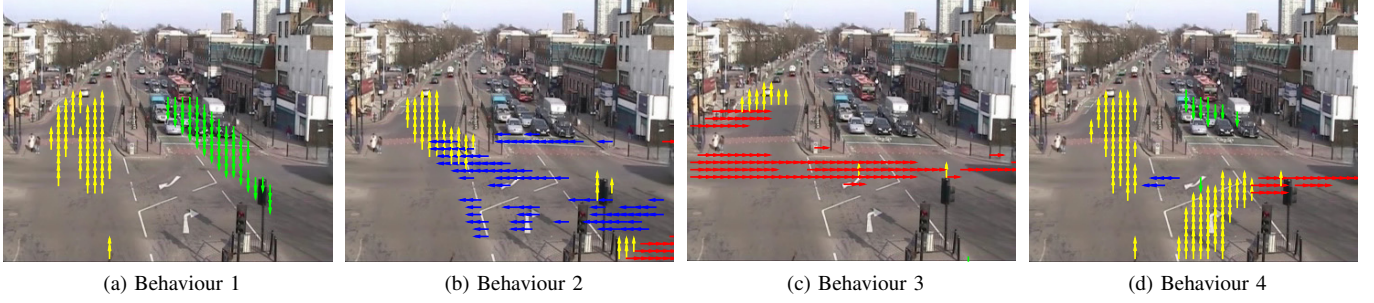


Figure 5. Behaviours learnt by the EM learning algorithm for the QMUL data. The arrows represent the visual words: the location and direction of the motion. The first behaviour (a) corresponds to the vertical traffic flow, the second (b) and the third (c) behaviours correspond to the left and right traffic flow respectively. The fourth (d) behaviour correspond to turns that follow the vertical traffic flow.

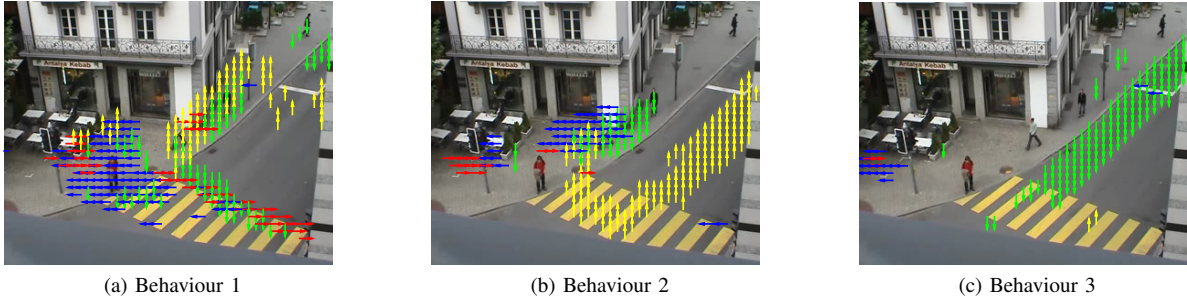


Figure 6. Behaviours learnt by the EM learning algorithm for the Idiap data. The arrows represent the visual words: the location and direction of the motion. The first behaviour (a) corresponds to the pedestrians motion, the second (b) and the third (c) behaviours correspond to the upward and downward traffic flows respectively.

(the first behaviour in Figure 6a), when cars stop in front of the stop line and pedestrians cross the road; a downward traffic flow (the third behaviour in Figure 6c), when cars move downward along the road; an upward traffic flow (the second behaviour in Figure 6b), when cars from left and right sides move upward on the road.

#### D. Anomaly detection

In this section the anomaly detection performance achieved by all three learning algorithms is compared. The datasets contain the number of abnormal events, such as jaywalking, car moving on the opposite lane, disruption of the traffic flow (see examples in Figure 7).

For the EM learning algorithm the plug-in approximation of the marginal likelihood is used for anomaly detection. For both the VB and GS learning algorithms both the plug-in and Monte Carlo approximations of the likelihood are used. Note, that for the GS algorithm samples are obtained during the learning stage. As 5 independent samples from the GS scheme are taken, the Monte Carlo approximation of the marginal likelihood is computed based on these 5 samples. For the VB learning algorithm samples are obtained after the learning stage from the posterior distributions, parameters of which are learnt. This means that the number of samples that are used for anomaly detection does not influence on the computational cost of learning. We test the Monte Carlo approximation of the marginal likelihood with 5 and 100 samples for the VB learning algorithm.

As a result, we have 18 methods to compare: obtained

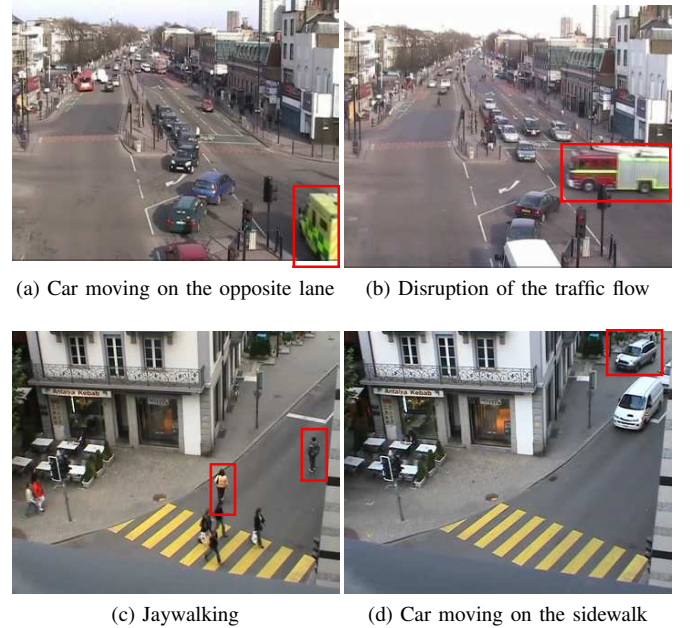


Figure 7. Examples of abnormal events

by three learning algorithms, three different groups of hyperparameters settings, one type of marginal likelihood approximation for the EM learning algorithm, two types of marginal likelihood approximation for the VB and GS learning algorithms, where for the former there are two Monte Carlo approximations using 5 and 100 samples. The list of methods

Table I  
METHODS REFERENCES

Reference	Learning algorithm	Hyper-parameters settings	Marginal likelihood approximation	Number of posterior samples
EM 1 p	EM	type 1	Plug-in	—
EM H p	EM	type H	Plug-in	—
EM H+1 p	EM	type H+1	Plug-in	—
VB 1 p	VB	type 1	Plug-in	—
VB 1 mc 5	VB	type 1	Monte Carlo	5
VB 1 mc 100	VB	type 1	Monte Carlo	100
VB H p	VB	type H	Plug-in	—
VB H mc 5	VB	type H	Monte Carlo	5
VB H mc 100	VB	type H	Monte Carlo	100
VB H+1 p	VB	type H+1	Plug-in	—
VB H+1 mc 5	VB	type H+1	Monte Carlo	5
VB H+1 mc 100	VB	type H+1	Monte Carlo	100
GS 1 p	GS	type 1	Plug-in	—
GS 1 mc	GS	type 1	Monte Carlo	5
GS H p	GS	type H	Plug-in	—
GS H mc	GS	type H	Monte Carlo	5
GS H+1 p	GS	type H+1	Plug-in	—
GS H+1 mc	GS	type H+1	Monte Carlo	5

references can be found in Table I.

Note, that we achieve a very fast decision making performance in our framework. Indeed, anomaly detection is made for approximately 0.0044 sec per visual document by the plug-in approximation of the marginal likelihood, for 0.0177 sec per document by the Monte Carlo approximation with 5 samples and for 0.3331 sec per document by the Monte Carlo approximation with 100 samples<sup>2</sup>.

The mean areas under precision-recall curves for anomaly detection for all 18 compared methods can be found in Figure 8. Below we examine the results with respect to hyperparameters sensitivity, an influence of the likelihood approximation on the final performance, we also compare the learning algorithms and discuss anomaly localisation results.

1) *Hyperparameters sensitivity*: This section presents sensitivity analysis of the anomaly detection methods with respect to changes of the hyperparameters.

The analysis of the mean areas under curves (Figure 8) suggests that the hyperparameters almost do not influence on the results of the EM learning algorithm, while there is significant dependence of hyperparameters changes and results of the VB and GS learning algorithms. These conclusions are confirmed by examination of the individual runs of the algorithms. For example, Figure 9 presents the precision-recall curves for all 20 runs with different initialisations of 4 methods for the Idiap data: the VB learning algorithm using the plug-in approximation of the marginal likelihood with the prior types 1 and H of the hyperparameters settings and the EM learning algorithm with the same prior groups of the hyperparameters settings. One can notice that the variance of the curves for the VB learning algorithm with the prior type 1 is larger than the corresponding variance with the prior type H, while the

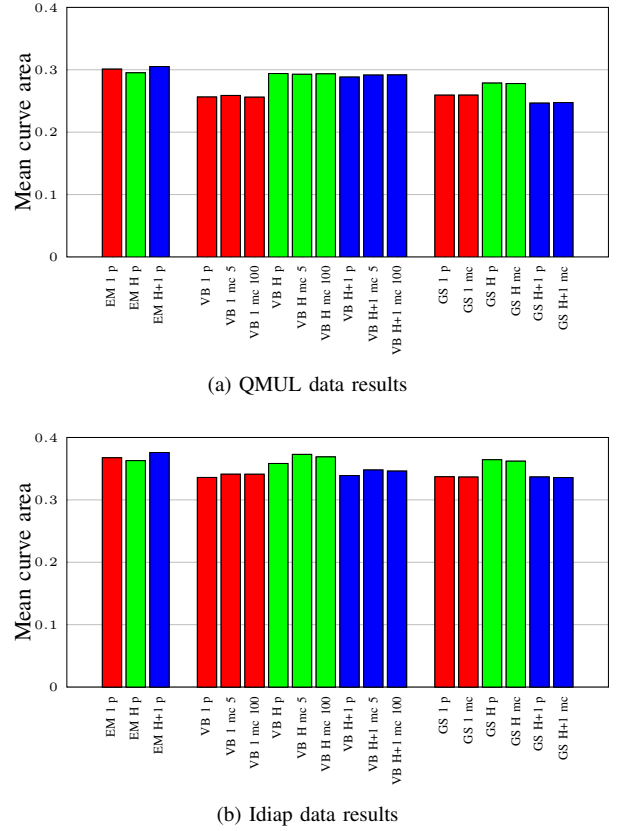


Figure 8. Results of anomaly detection. (a) are the mean areas under precision-recall curves for the QMUL data. (b) are the mean areas under precision-recall curves for the Idiap data.

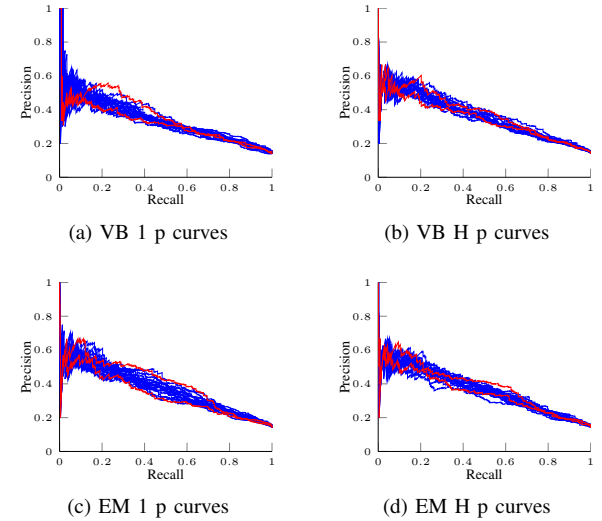


Figure 9. Hyperparameters sensitivity of the precision-recall curves. The top row corresponds to all the independent runs of the VB learning algorithm with the prior type 1 (a) and the prior type H (b). The bottom row corresponds to all the independent runs of the EM learning algorithm with the prior type 1 (c) and the prior type H (d). The red colour highlights the curves with the maximum and minimum areas under curves.

similar variances for the EM learning algorithm are very close to each other.

Note, that the results of the EM learning algorithm with the prior type 1 do not significantly differ from the results with the

<sup>2</sup>The computational time is provided for a laptop computer with i7-4720HQ CPU with 2.20GHz, 16 GB RAM using Matlab R2015a implementation.



Table II  
MEAN AREA UNDER PRECISION-RECALL CURVES

Dataset	EM	VB	GS
QMUL	0.3051	0.2940	0.2787
Idiap	0.3759	0.3729	0.3643

other priors, despite of the fact that the prior type 1 actually cancels out the prior influence on the parameters estimates and equates the MAP and maximum likelihood estimates. We can conclude that the choice of the hyperparameters settings is not a problem for the EM learning algorithm and we can even simplify the derivations considering only the maximum likelihood estimates without the prior influence.

The VB and GS learning algorithms require a proper choice of the hyperparameters settings as they can significantly change the anomaly detection performance. This choice can be performed empirically or with the type II maximum likelihood approach [35].

2) *Marginal likelihood approximation influence:* In this section the influence of the type of the marginal likelihood approximation on the anomaly detection results is studied.

The average results for both datasets (Figure 8) demonstrate that the type of the marginal likelihood approximation does not influence remarkably on anomaly detection performance. As the plug-in approximation requires less computational resources both in terms of time and memory (as there is no need to sample and store posterior samples and average among them) this type of approximation is recommended to be used for anomaly detection in the proposed framework.

3) *Learning algorithms comparison:* This section compares the anomaly detection performance obtained by three learning algorithms.

The best results in terms of a mean area under a precision-recall curve are obtained by the EM learning algorithm, the worst results are obtained by the GS learning algorithm (Figure 8 and Table II). In Table II for each learning algorithm the group of hyperparameters settings and the type of marginal likelihood approximation is chosen to have the maximum of the mean area under curves, where a mean is taken over independent runs of the same method and maximum is taken among different settings for the same learning algorithm.

Figure 10 presents the best and the worst precision-recall curves (in terms of the area under them) for the individual runs of the learning algorithms. The figure shows that among the individual runs the EM learning algorithm also demonstrates the most accurate results. Although, the minimum area under the precision-recall curve for the EM learning algorithm is less than the area under the corresponding curve for the VB algorithm. It means that the variance among the individual curves for the EM learning algorithm is larger in comparison with the VB learning algorithm.

The variance of the precision-recall curves for both VB and GS learning algorithms are relatively small. However, the VB learning algorithm has the curves higher than the curves obtained by the GS learning algorithm. It can be confirmed by examination of the best and worst precision-recall curves (Figure 10) and the mean values of the area under curves

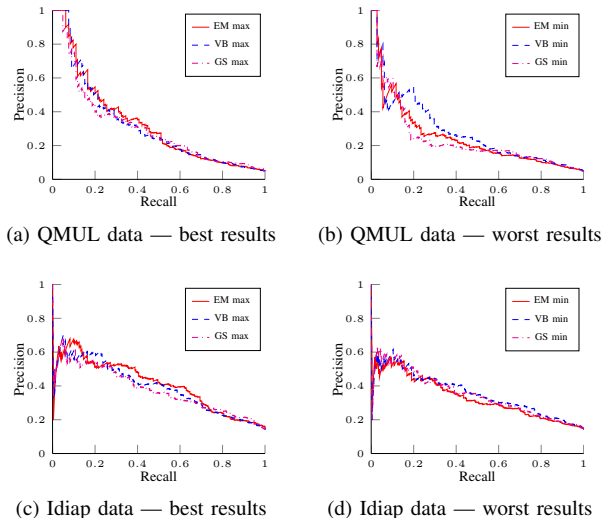


Figure 10. Precision-recall curves with the maximum and minimum areas under curves for the three learning algorithms (maximum and minimum is among all the runs with different initialisations for all groups of hyperparameters settings and all types of marginal likelihood approximations). (a) presents the “best” curves for the QMUL data, i.e. the curves with the maximum area under a curve. (b) presents the “worst” curves for the QMUL data, i.e. the curves with the minimum area under a curve. (c) presents the “best” curves for the Idiap data, (d) — the “worst” curves for the Idiap data.

Table III  
BEST CLASSIFICATION ACCURACY FOR THE EM LEARNING ALGORITHM

Dataset	Accuracy
QMUL	0.9544
Idiap	0.8891

(Figure 8 and Table II).

We also present the results of classification accuracy, i.e., the fraction of the correctly classified documents, for anomaly detection, which can be achieved with some fixed threshold. The best classification accuracy for the EM learning algorithm in both datasets can be found in Table III.

4) *Anomaly localisation:* We apply the proposed method for anomaly localisation, presented in Section V-B, and get promising results. We demonstrate the localisation results for the EM learning algorithm with the prior type H+1 on both datasets in Figure 11. The red rectangle is manually set to locate the abnormal events within the frame, the arrows correspond to the visual words with the smallest marginal likelihood computed by the algorithm. It can be seen that the abnormal events correctly localised by the proposed method.

## VII. CONCLUSIONS

This paper presents two learning algorithms for the dynamic topic model for behaviour analysis in video: the EM-algorithm is developed for the MAP estimates of the model parameters and a variational Bayes inference algorithm to developed for calculating the posterior distributions of them. A detailed comparison of these proposed learning algorithms with the Gibbs sampling based algorithm developed in [23] is presented. The differences and the similarities of the theoretical aspects for all three learning algorithms are well emphasised. The empirical



Figure 11. Examples of anomalies localisation. The red rectangle is the manual localisation. The arrows represent the visual words with the smallest marginal likelihood, the locations of the arrows are the results of the algorithmic anomalies localisation.

comparison is performed for abnormal behaviour detection using two unlabelled real video datasets. Both proposed learning algorithms demonstrate more accurate results than the algorithm proposed in [23] in terms of anomaly detection performance.

The EM learning algorithm demonstrates the best results in terms of the mean values of the performance measure, obtained by the independent runs of the algorithm with different random initialisations. Although, it is noticed that the variance among the precision-recall curves of the individual runs is relatively high. The variational Bayes learning algorithm shows the smaller variance among the precision-recall curves than the EM-algorithm. The results show that the VB algorithm answers are more robust to different initialisation values. However, it is shown that the results of the algorithm are significantly influenced by the choice of the hyperparameters. The hyperparameters require additional tuning before the algorithm can be applied to data. Note, that the results of the EM learning algorithm only slightly depend on the choice of the hyperparameters settings. Moreover, the hyperparameters can be even set in such a way as the EM algorithm is applied to obtain the maximum likelihood estimates instead of the maximum a posteriori ones. Both proposed learning algorithms — EM and VB — provide more accurate results in comparison to the Gibbs sampling based algorithm.

We also demonstrate that consideration of marginal likelihoods of visual words rather than visual documents can provide satisfactory results about locations of anomalies within a frame. In our best knowledge the proposed localisation procedure is the first general approach in probabilistic topic modeling that requires only presence of spatial information encoded in visual words.

## APPENDIX A EM-ALGORITHM DERIVATIONS

This Appendix presents the details of the proposed EM learning algorithm derivation. The objective function in the EM-algorithm is:

$$\begin{aligned}
 \mathcal{Q}(\Omega, \Omega^{old}) + \log p(\Omega | \beta, \alpha, \eta, \gamma) = & \sum_{\mathbf{y}_{1:T}} \sum_{\mathbf{z}_{1:T}} \left( p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T} | \mathbf{x}_{1:T}, \Omega^{old}) \times \right. \\
 & \left. \log p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}, \mathbf{z}_{1:T} | \Omega, \alpha, \beta, \gamma, \eta) \right) + \\
 & + \log p(\Omega | \beta, \alpha, \eta, \gamma) = \\
 = Const + \sum_{z_1 \in \mathcal{Z}} \left( \log \pi_{z_1} p(z_1 | \mathbf{x}_{1:T}, \Omega^{old}) \right) + & \\
 \sum_{t=2}^T \sum_{z_t \in \mathcal{Z}} \sum_{z_{t-1} \in \mathcal{Z}} \left( \log \xi_{z_t, z_{t-1}} p(z_t, z_{t-1} | \mathbf{x}_{1:T}, \Omega^{old}) \right) + & \\
 \sum_{t=1}^T \sum_{i=1}^{N_t} \sum_{y_{i,t} \in \mathcal{Y}} \left( \log \phi_{x_{i,t}, y_{i,t}} p(y_{i,t} | \mathbf{x}_{1:T}, \Omega^{old}) \right) + & \\
 + \sum_{t=1}^T \sum_{i=1}^{N_t} \sum_{z_t \in \mathcal{Z}} \sum_{y_{i,t} \in \mathcal{Y}} \left( \log \theta_{y_{i,t}, z_t} p(y_{i,t}, z_t | \mathbf{x}_{1:T}, \Omega^{old}) \right) + & \\
 \sum_{z \in \mathcal{Z}} (\eta_z - 1) \log \pi_z + \sum_{z \in \mathcal{Z}} \sum_{z' \in \mathcal{Z}} (\gamma_z - 1) \log \xi_{z, z'} + & \\
 \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{Y}} (\alpha_y - 1) \log \theta_{y, z} + \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} (\beta_x - 1) \log \phi_{x, y} & \quad (44)
 \end{aligned}$$

On the M-step the function (44) is maximised with respect to the parameters  $\Omega$  with fixed values for  $p(z_1 | \mathbf{x}_{1:T}, \Omega^{old})$ ,  $p(z_t, z_{t-1} | \mathbf{x}_{1:T}, \Omega^{old})$ ,  $p(y_{i,t} | \mathbf{x}_{1:T}, \Omega^{old})$ ,  $p(y_{i,t}, z_t | \mathbf{x}_{1:T}, \Omega^{old})$ . The optimisation problem can be solved separately for each parameter, which leads to the equations (6) – (8).

On the E-step for the efficient implementation the forward-backward steps are developed for the auxiliary variables  $\hat{\alpha}_z(t)$  and  $\hat{\beta}_z(t)$ :

$$\begin{aligned}
 \hat{\alpha}_z(t) \stackrel{\text{def}}{=} p(\mathbf{x}_1, \dots, \mathbf{x}_t, z_t = z | \Omega^{old}) = & \\
 \sum_{\mathbf{z}_{1:t-1}} \pi_{z_1}^{old} \left[ \prod_{t=2}^{t-1} \xi_{z_t, z_{t-1}}^{old} \right] \left[ \prod_{t=1}^{t-1} \sum_{i=1}^{N_t} \sum_{y \in \mathcal{Y}} \phi_{x_{i,t}, y}^{old} \theta_{y, z_t}^{old} \right] \times & \\
 \xi_{z_t=k, z_{t-1}}^{old} \prod_{i=1}^{N_t} \sum_{y \in \mathcal{Y}} \phi_{x_{i,t}, y}^{old} \theta_{y, z_t=k}^{old} & \quad (45)
 \end{aligned}$$

Reorganisation of the terms in (45) leads to the recursive expressions (10).

Similarly for  $\hat{\beta}_z(t)$ :

$$\begin{aligned}
 \hat{\beta}_k(t) \stackrel{\text{def}}{=} p(\mathbf{x}_{t+1}, \dots, \mathbf{x}_T | z_t = z, \Omega^{old}) = & \\
 \sum_{\mathbf{z}_{t+1:T}} \xi_{z_{t+1}, z_t=k}^{old} \left[ \prod_{t=t+2}^T \xi_{z_t, z_{t-1}}^{old} \right] \prod_{t=t+1}^T \sum_{i=1}^{N_t} \sum_{y \in \mathcal{Y}} \phi_{x_{i,t}, y}^{old} \theta_{y, z_t=k}^{old} & \quad (46)
 \end{aligned}$$

The recursive formula (11) is obtained by interchanging the terms in (46).

The required posterior of the hidden variables terms  $p(z_1|\mathbf{x}_{1:T}, \mathbf{\Omega}^{Old})$ ,  $p(z_t, z_{t-1}|\mathbf{x}_{1:T}, \mathbf{\Omega}^{Old})$ ,  $p(y_{i,t}|\mathbf{x}_{1:T}, \mathbf{\Omega}^{Old})$ ,  $p(y_{i,t}, z_t|\mathbf{x}_{1:T}, \mathbf{\Omega}^{Old})$  are then expressed via the axillary variables  $\hat{\alpha}_z(t)$  and  $\hat{\beta}_z(t)$ , which leads to (13) – (16).

## APPENDIX B VB ALGORITHM DERIVATIONS

This Appendix presents the details of the proposed variational Bayes inference derivation. We have separated the parameters and the hidden variables. Let us consider the update formula of the variational Bayes inference scheme [32] for the parameters:

$$\begin{aligned} \log q(\mathbf{\Omega}) = Const + \\ \mathbb{E}_{q(\mathbf{y}_{1:T}, \mathbf{z}_{1:T})} \log p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \mathbf{\Omega} | \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \\ Const + \mathbb{E}_{q(\mathbf{y}_{1:T}, \mathbf{z}_{1:T})} \left( \sum_{z \in \mathcal{Z}} (\eta_z - 1) \log \pi_z + \right. \\ \sum_{z \in \mathcal{Z}} \sum_{\tilde{z} \in \mathcal{Z}} (\gamma_{\tilde{z}} - 1) \log \xi_{\tilde{z}, z} + \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{Y}} (\alpha_y - 1) \log \theta_{y, z} + \\ \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} (\beta_x - 1) \log \phi_{x, y} + \sum_{z \in \mathcal{Z}} \mathbb{I}(z_1 = z) \log \pi_z + \\ \sum_{t=2}^T \sum_{z \in \mathcal{Z}} \sum_{\tilde{z} \in \mathcal{Z}} \mathbb{I}(z_t = \tilde{z}) \mathbb{I}(z_{t-1} = z) \log \xi_{\tilde{z}, z} + \\ \sum_{t=1}^T \sum_{i=1}^{N_t} \sum_{y \in \mathcal{Y}} \mathbb{I}(y_{i,t} = y) \log \phi_{x_{i,t}, y} + \\ \left. \sum_{t=1}^T \sum_{i=1}^{N_t} \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{Y}} \mathbb{I}(y_{i,t} = y) \mathbb{I}(z_t = z) \log \theta_{y, z} \right) \quad (47) \end{aligned}$$

One can notice that  $\log q(\mathbf{\Omega})$  is further factorised as in (18). Now each factorisation term can be considered independently. Derivations of the equations (19) – (22) are very similar to each other. We provide the derivation only of the term  $q(\Phi)$ :

$$\begin{aligned} \log q(\Phi) = Const + \\ \mathbb{E}_{q(\mathbf{y}_{1:T}, \mathbf{z}_{1:T})} \left( \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} (\beta_x - 1) \log \phi_{x, y} + \right. \\ \left. \sum_{t=1}^T \sum_{i=1}^{N_t} \sum_{y \in \mathcal{Y}} \mathbb{I}(y_{i,t} = y) \log \phi_{x_{i,t}, y} \right) = \\ Const + \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} (\beta_x - 1) \log \phi_{x, y} + \\ \sum_{t=1}^T \sum_{i=1}^{N_t} \sum_{y \in \mathcal{Y}} \log \phi_{x_{i,t}, y} \underbrace{\mathbb{E}_{q(\mathbf{y}_{1:T}, \mathbf{z}_{1:T})} (\mathbb{I}(y_{i,t} = y))}_{q(y_{i,t}=y)} = \\ Const + \\ \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} \log \phi_{x, y} \left( \beta_x - 1 + \sum_{t=1}^T \sum_{i=1}^{N_t} \mathbb{I}(x_{i,t} = x) q(y_{i,t} = y) \right) \quad (48) \end{aligned}$$

It can be noticed from (48) that the distribution of  $\Phi$  is a product of the Dirichlet distributions (19).

The update formula in the variational Bayes inference scheme for the hidden variables is as follows:

$$\begin{aligned} \log q(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}) = Const + \\ \mathbb{E}_{q(\boldsymbol{\pi})q(\Xi)q(\Theta)q(\Phi)} \log p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}, \mathbf{z}_{1:T}, \mathbf{\Omega} | \boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \\ Const + \sum_{z \in \mathcal{Z}} \mathbb{I}(z_1 = z) \mathbb{E}_{q(\boldsymbol{\pi})} \log \pi_z + \\ \sum_{t=2}^T \sum_{z \in \mathcal{Z}} \sum_{\tilde{z} \in \mathcal{Z}} \mathbb{I}(z_t = \tilde{z}) \mathbb{I}(z_{t-1} = z) \mathbb{E}_{q(\Xi)} \log \xi_{\tilde{z}, z} + \\ \sum_{t=1}^T \sum_{i=1}^{N_t} \sum_{y \in \mathcal{Y}} \mathbb{I}(y_{i,t} = y) \mathbb{E}_{q(\Phi)} \log \phi_{x_{i,t}, y} + \\ \sum_{t=1}^T \sum_{i=1}^{N_t} \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{Y}} \mathbb{I}(y_{i,t} = y) \mathbb{I}(z_t = z) \mathbb{E}_{q(\Theta)} \log \theta_{y, z} \quad (49) \end{aligned}$$

We know from the parameters update (19) – (22) that their distributions are Dirichlet. Therefore  $\mathbb{E}_{q(\boldsymbol{\pi})} \log \pi_z = \psi(\tilde{\eta}_z) - \psi(\sum_{z' \in \mathcal{Z}} \tilde{\eta}_{z'})$  (see, for example, [32]) and similarly for all the other expected value expressions.

Using the introduced notations (23) – (26) the update formula (49) for the hidden variables can be then expressed as:

$$\begin{aligned} \log q(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}) = Const + \sum_{z \in \mathcal{Z}} \mathbb{I}(z_1 = z) \log \tilde{\pi}_z + \\ \sum_{t=2}^T \sum_{z \in \mathcal{Z}} \sum_{\tilde{z} \in \mathcal{Z}} \mathbb{I}(z_t = \tilde{z}) \mathbb{I}(z_{t-1} = z) \log \tilde{\xi}_{\tilde{z}, z} + \\ \sum_{t=1}^T \sum_{i=1}^{N_t} \sum_{y \in \mathcal{Y}} \mathbb{I}(y_{i,t} = y) \log \tilde{\phi}_{x_{i,t}, y} + \\ \sum_{t=1}^T \sum_{i=1}^{N_t} \sum_{z \in \mathcal{Z}} \sum_{y \in \mathcal{Y}} \mathbb{I}(y_{i,t} = y) \mathbb{I}(z_t = z) \log \tilde{\theta}_{y, z} \quad (50) \end{aligned}$$

The approximated distribution of the hidden variables is then:

$$\begin{aligned} q(\mathbf{y}_{1:T}, \mathbf{z}_{1:T}) = \\ \frac{1}{\tilde{K}} \tilde{\pi}_{z_1} \left[ \prod_{t=2}^T \tilde{\xi}_{z_t, z_{t-1}} \right] \prod_{t=1}^T \prod_{i=1}^{N_t} \tilde{\phi}_{x_{i,t}, y_{i,t}} \tilde{\theta}_{y_{i,t}, z_t}, \quad (51) \end{aligned}$$

where  $\tilde{K}$  is a normalisation constant. Note that the expression of the true posterior distribution of the hidden variables is the same up to replacing the true parameters variables with the corresponding tilde variables:

$$\begin{aligned} p(\mathbf{y}_{1:T}, \mathbf{z}_{1:T} | \mathbf{x}_{1:T}, \mathbf{\Omega}) = \\ \frac{1}{K} \pi_{z_1} \left[ \prod_{t=2}^T \xi_{z_t, z_{t-1}} \right] \prod_{t=1}^T \prod_{i=1}^{N_t} \phi_{x_{i,t}, y_{i,t}} \theta_{y_{i,t}, z_t} \quad (52) \end{aligned}$$

Therefore to compute the required expressions of the hidden variables  $q(z_1 = z)$ ,  $q(z_{t-1} = z, z_t = z')$ ,  $q(y_{i,t} = y, z_t = z)$  and  $q(y_{i,t} = y)$  one can use the same forward-backward procedure and update formula as in the E-step of the EM-algorithm replacing all the parameters variables with the corresponding introduced tilde variables.

## ACKNOWLEDGMENTS

Olga Isupova and Lyudmila Mihaylova would like to thank the support from the EC Seventh Framework Programme [FP7 2013-2017] TRacking in complex sensor systems (TRAX) Grant agreement no.: 607400. Lyudmila Mihaylova also acknowledges the UK Engineering and Physical Sciences Research Council (EPSRC) for the support via the Bayesian Tracking and Reasoning over Time (BTaRoT) grant EP/K021516/1.

## REFERENCES

- [1] R. Raghuvaran, A. Del Bue, M. Cristani, and V. Murino, "Optimizing interaction force for global anomaly detection in crowded scenes," in *Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, November 2011, pp. 136–143.
- [2] M. Roshkhar and M. Levine, "Online dominant and anomalous behavior detection in videos," in *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 2611–2618.
- [3] S.-H. Yen and C.-H. Wang, "Abnormal event detection using HOSF," in *Proceedings of the International Conference on IT Convergence and Security (ICITCS) 2013*. IEEE, 2013, pp. 1–4.
- [4] K. Ouyirach, S. Gharti, and M. N. Dailey, "Incremental behavior modeling and suspicious activity detection," *Pattern Recognition*, vol. 46, no. 3, pp. 671 – 680, 2013.
- [5] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information theory*, vol. 37, pp. 145–151, 1991.
- [6] Z. Su, H. Wei, and S. Wei, "Crowd event perception based on spatiotemporal Weber field," *Journal of Electronical and Computer Engineering*, vol. 2014, Jan. 2014.
- [7] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Abnormal crowd behavior detection and localization using maximum sub-sequence search," in *Proceedings of the 4th ACM/IEEE International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream*, ser. ARTEMIS '13. New York, NY, USA: ACM, 2013, pp. 49–58.
- [8] Y.-H. Liu, Y.-C. Liu, and Y.-J. Chen, "Fast support vector data descriptions for novelty detection," *IEEE Transactions on Neural Networks*, vol. 21, no. 8, pp. 1296–1313, August 2010.
- [9] L. Maddalena and A. Petrosino, "Stopped object detection by learning foreground model in videos," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 5, pp. 723–735, May 2013.
- [10] S. Osher and J. A. Sethian, "Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations," *Journal of Computational Physics*, vol. 79, no. 1, pp. 12–49, 1988.
- [11] X. Ding, Y. Li, A. Belatreche, and L. Maguire, "Novelty detection using level set methods," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 3, pp. 576–588, March 2015.
- [12] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, March 2008.
- [13] A. Feizi, A. Aghagholzadeh, and H. Seyedarabi, "Using optical flow and spectral clustering for behavior recognition and detection of anomalous behaviors," in *Proceedings of the 8th Iranian Conference on Machine Vision and Image Processing (MVIP) 2013*, September 2013, pp. 210–213.
- [14] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, June 2008, pp. 1–8.
- [15] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.*, June 2009, pp. 1446–1453.
- [16] C. Brighenti and M. Sanz-Bobi, "Auto-regressive processes explained by self-organized maps. application to the detection of abnormal behavior in industrial processes," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2078–2090, December 2011.
- [17] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 50–57.
- [18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
- [19] R. Mehrotra, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 935–942.
- [20] J. Li, S. Gong, and T. Xiang, "Global behaviour inference using probabilistic latent semantic analysis," in *Proceeding of the British Machine Vision Conference*, 2008, pp. 193–202.
- [21] J. Varadarajan and J. Odobez, "Topic models for scene analysis and abnormality detection," in *Proceedings of the 12th IEEE International Conference on Computer Vision Workshops (ICCV Workshops) 2009*, September 2009, pp. 1338–1345.
- [22] X. Wang and X. Ma, "Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 539 – 555, 2009.
- [23] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *International Journal of Computer Vision*, vol. 98, no. 3, pp. 303–323, 2012.
- [24] J. Varadarajan, R. Emonet, and J.-M. Odobez, "A sparsity constraint for topic models - application to temporal activity mining," in *Proceedings of the NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions*, December 2010.
- [25] D. Kuettel, M. Breitenstein, L. Van Gool, and V. Ferrari, "What's going on? Discovering spatio-temporal dependencies in dynamic scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*, June 2010, pp. 1951–1958.
- [26] H. Jeong, Y. Yoo, K. M. Yi, and J. Y. Choi, "Two-stage online inference model for traffic pattern analysis and anomaly detection," *Machine Vision and Applications*, vol. 25, no. 6, pp. 1501–1517, 2014.
- [27] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.
- [28] J.-T. Chien and M.-S. Wu, "Adaptive Bayesian latent semantic analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 198–207, Jan 2008.
- [29] O. Isupova, L. Mihaylova, D. Kuzin, G. Markarian, and F. Septier, "An expectation maximisation algorithm for behaviour analysis in video," in *Proceeding of the 18th International Conference on Information Fusion (Fusion) 2015*, July 2015, pp. 126–133.
- [30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.
- [31] K. Vorontsov and A. Potapenko, "Additive regularization of topic models," *Machine Learning*, pp. 1–21, 2014.
- [32] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
- [33] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.
- [34] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, "On smoothing and inference for topic models," in *Proceedings of the 25-th Conference on Uncertainty in Artificial Intelligence*, ser. UAI '09. Arlington, Virginia, United States: AUAI Press, 2009, pp. 27–34.
- [35] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [36] T. S. F. Haines and T. Xiang, "Video topic modelling with behavioural segmentation," in *Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis*, ser. MPVA '10. New York, NY, USA: ACM, 2010, pp. 53–58.
- [37] D. Pathak, A. Sharang, and A. Mukerjee, "Anomaly localization in topic-based analysis of surveillance videos," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision 2015*, Jan 2015, pp. 389–395.